Eficiencia de Google Noticias en la recopilación de archivos hemerográficos

Crúz Mena, Javier, García Rojo María Keninseb, Alvarado Cruz Isela.

Palabras clave:

periodismo de ciencia, motor de búsqueda, Google, análisis de contenido, consultas simples

Introducción

El ejercicio del periodismo de ciencia va más allá de proveer información sobre avances científicos o tecnológicos. Parte importante de este quehacer consiste en que el periodista sea capaz de dotar a sus productos periodísticos de información y conocimientos básicos sobre el tema para facilitar a los ciudadanos la toma de decisiones respecto de su propia persona y de su colectividad, como lo explican algunos autores.[i] De ahí la importancia de evaluar el contenido de la prensa escrita a partir de parámetros tales como el carácter noticioso del hecho, las fuentes de información consultadas y la forma en que se presentan al lector.

Para hacer esta tarea, comúnmente, la hemeroteca es la primera fuente a la que se acude cuando deseamos revisar o recopilar notas periodísticas atrasadas. El mecanismo de búsqueda se efectúa de forma manual y una revisión minuciosa de los ejemplares puede garantizar un rastreo completo de las notas. No obstante, los obstáculos a los que se enfrenta el usuario durante la recopilación, como la demora en la entrega del material solicitado; material hemerográfico en restauración; falta de ejemplares en el archivo; material en posesión de otro usuario; material incompleto y sobre todo saber, de antemano, que debido a su impacto social el número de notas del hecho será amplio, esto convierte a

la búsqueda, al final, en tiempo significativamente perdido. Por esta razón, nos propusimos usar un mecanismo que cumpliera con criterios determinados y redujera el tiempo empleado en la recolección del material periodístico.

Trejo Delarbre (1996) argumenta que la digitalización de la información es el sustento de la nueva revolución informática. La singularidad del Internet es la facilidad para que diversos grupos cuenten con acceso no sólo a la recepción de mensajes, sino también a la propagación de ellos. Por otra parte, Islas (2005) agrega que la Internet posee la capacidad de proporcionar servicios personalizados que respondan a las exigencias de cada usuario.

A partir de este cúmulo de ventajas cibernéticas, decidimos buscar en la red de redes una herramienta de rastreo que cumpliera nuestras demandas de búsqueda de archivos hemerográficos de origen diverso.

Hoy en día los motores de búsqueda son el mecanismo primario para encontrar información en la red.[ii] Son considerados "una aplicación de Internet utilizados para localizar documentos y páginas web, partiendo de las *keywords* o palabras clave. Los motores de búsqueda más poderosos o sofisticados rastrean Internet en busca de sitios web y sus bases de datos para responder con el máximo de posibilidades a sus usuarios. Esta última acción la desarrollan los robots conocidos como *bots*, arañas y rastreadores, -en inglés spiders y *crawlers* respectivamente-. Una vez realizada la operación de búsqueda se ofrece en pantalla el listado de documentos con los enlaces que les corresponden para poder acceder tan solo haciendo clic en cualquiera de ellos".[iii]

Metodología

Partiendo del supuesto de que reunir el material por medio de los motores, a diferencia del método manual, aseguraría la reducción del tiempo empleado en esta tarea, consideramos indispensable, como primer paso, determinar criterios y homologar el mecanismo para una recopilación eficiente.

Tema	Reforma energética 2008	Influenza H1N1 2009
Palabra clave	pemex, energética	influenza, epidemia
Periodo	P1: 8-9 abril 2008 P2: 13-14 mayo 2008 P3: 23-24 octubre 2008 P4: 28-29 octubre 2008	P5: 23-24 abril 2009 P6: 29-30 abril 2009 P7: 4-5 mayo 2009 P8: 10-11 mayo 2009
Medios	www.jornada.unam.mx www.eluniversal.com.mx www.reforma.com	

FIGURA 1. Criterios de búsqueda

La selección de los temas responde a la relevancia que tuvieron en las agendas noticiosas de la prensa mexicana; asimismo, se trata de asuntos cuyas bases se sostienen en conocimientos científicos y sus efectos aún repercuten en la actualidad. Para este ejercicio elegimos el debate de la reforma energética en 2008 y el brote del virus de influenza A (H1N1) en 2009.

Los términos “energética y pemex” e “influenza y epidemia” fueron las palabras clave seleccionadas para este experimento, ya que engloban y representan el tema en general. Asimismo decidimos usar términos simples, los cuales incrementan la precisión en la

búsqueda, a diferencia de términos compuestos formados por dos o más palabras.[iv]

El periodo establecido para este ejercicio abarcó ocho periodos de dos días cada uno, en total 16 consultas. Para los primeros cuatro periodos se utilizaron las palabras clave "energética" y "pemex"; y para los otros cuatro se asignó "influenza" y "epidemia". Cada intervalo de fechas se ajusta al tiempo en que sucedieron los acontecimientos y su elección fue aleatoria.

Los medios o sitios en los cuales se efectuaron las consultas fueron: *La Jornada (*LJ), *Reforma* (R) y *El Universal* (U), debido a su impacto en la sociedad y a su tiraje.

Ya establecidos estos criterios, acudimos a la Hemeroteca Nacional de México para revisar cada ejemplar de manera exhaustiva y reunir las notas publicadas de cada diario; la revisión se efectuó en dos ocasiones para extraer el material completo y evitar la omisión de alguna nota. Posteriormente, seleccionamos entre los diversos motores de búsqueda en la Internet, aquél que cumpliera con nuestras necesidades.

Descubrimos que varios periódicos con formato electrónico ofrecen un buscador interno para rastrear información en sus propias páginas. No obstante, nos resistimos a utilizar este mecanismo a fin de evitar las diferencias e incluso las inconsistencias que los distintos motores podrían presentar y decidimos servirnos de una misma herramienta para los tres medios elegidos.

Durante el análisis elaboramos una lista con 29 motores de búsqueda, clasificados de acuerdo a las modalidades que ofrecían: Búsqueda Simple (BS)[v] y Búsqueda Avanzada (BA), siendo ésta última el primer parámetro para elegirlos. A diferencia de la primera, la BA

ofrece cuadros de texto para delimitar información en una consulta especializada y obtener resultados potencialmente mejores.

De los 29 sólo 19 motores contaron con Búsqueda Avanzada; no obstante, en la mayoría de los 19 resultados de BA se omitía la opción para definir el intervalo de fechas, lo cual obligaba a someter los resultados del experimento a un nuevo filtro tomando como referencia los tres campos específicos de la búsqueda: palabra clave, medio o sitio e intervalo de fechas.

Tras un análisis detallado, *Google*[vi] fue el único motor de búsqueda que ofreció esta refinación en las consultas. La BA en *Google* también carecía de uno de los parámetros solicitados, sin embargo en una revisión minuciosa encontramos que *Google* cuenta con una herramienta especial para extraer noticias: *La búsqueda Avanzada en el Archivo de Google Noticias*.

No	Motores de búsqueda	Dirección	Formato de búsqueda
1	CompuServe	(http://webcenters.netscape.compuserve.com/menu)	BS
2	Mamma	(http://www.mamma.com)	BS
3	Lycos	(http://search.lycos.com)	BS
4	Ciao	(http://www.ciao.es/teoma_com__330723)	BS
5	Scoopler	(http://www.scoopler.com)	BS
6	Onriot	(http://www.oneriot.com)	BS
7	Todalanet	(http://www.todalanet.net/)	BS
8	Ipselon	(http://ipselon.com/es/)	BS

9	About	(http://www.about.com/)	BS
10	Webcrawler	(http://www.webcrawler.com/)	BS y la búsqueda la realiza a través de otros motores: *Google, Yahoo, Bing y Ask*
11	MSN	(http://prodigy.msn.com)	BS y BA. En la BA no permite elegir un intervalo de fechas y la realiza a través de *Bing*
12	AOL	(http://www.aol.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
13	Ask	(http://es.ask.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
14	Altavista	(http://mx.altavista.com)	BS y BA. La BA cuenta con los elementos requeridos.[vii]
15	Gigablast	(http://www.gigablast.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
16	Snap	(http://www.snap.com)	El formato de BS y BA es el mismo
17	Yahoo	(http://mx.yahoo.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
18	Alltheweb	(http://www.alltheweb.com)	BS y BA. En BA solo falta medio
19	Hotbot	(http://www.hotbot.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
20	Infospace	(http://search.infospace.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
21	Metacrawler	(http://www.metacrawler.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
22	Ixquick	(http://www.ixquick.com/esp/)	BS y BA. En la BA no permite elegir un intervalo de fechas
23	Dogpile	(http://www.dogpile.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
24	Search	(http://www.search.com)	BS y BA. En la BA no permite elegir un intervalo de fechas
25	Mostercrawler	(http://monstercrawler.com/)	BS y BA. En la BA no permite elegir un intervalo de fechas.la realiza a través de otros motores: *Google, Yahoo, Bing y Ask*
26	Bing	(http://www.bing.com/)	BS y BA. En la BA no permite elegir un intervalo de fechas
27	Excite	(http://www.excite.com/)	BS y BA. En la BA no permite elegir un intervalo de fechas
28	Go	(http://www.go.com/)	BS y BA. En la BA no permite elegir un intervalo de fechas y la realiza a través de *Yahoo*

29	Google	(http://.www.google.com/)	BS y BA. La BA cuenta con los elementos requeridos.

FIGURA 2. Clasificación de los motores de búsqueda
La tabla muestra la clasificación en motores con Búsqueda Simple (BS) y Búsqueda Avanzada (BA), en esta última clasificación se describen las características con las que cuenta cada buscador.

Para acceder a la aplicación de *La búsqueda Avanzada en el Archivo de Google Noticias* usamos la dirección URL: **http://news.google.com.mx/archivesearch/advanced_search**, la cual nos despliega la siguiente ventana:

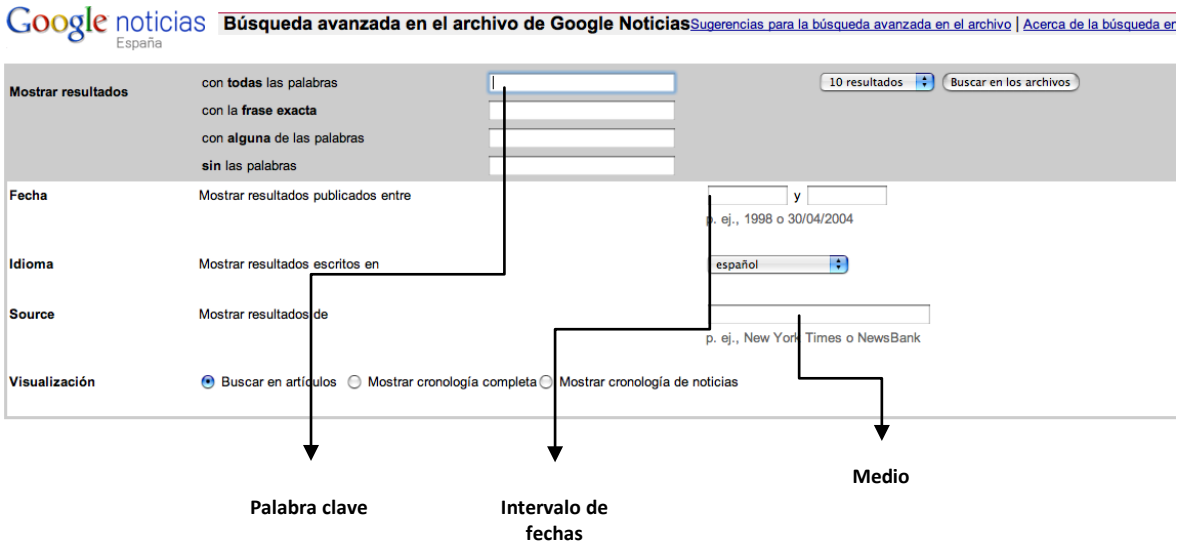


FIGURA 3. Ventana principal de *Búsqueda avanzada en el archivo de Google Noticias*
La ventana muestra los campos específicos para realizar la búsqueda con frase exacta, fecha, medio, etcétera

Resultados

De la revisión directa en la hemeroteca, elaboramos una tabla con los resultados de los ejemplares y descubrimos que el comportamiento de la cobertura hecha por los tres periódicos para cada tema es cuantitativamente similar: de la Reforma energética *La Jornada* obtuvo 218 notas; *Reforma* 195 y *El Universal* 128 notas; mientras que para el tema de la influenza A (H1N1) *La Jornada* muestra 466, *Reforma* 494 y *El Universal 429 notas*. Tras la revisión de los tres diarios en las fechas establecidas (16 pares de días, cuatro para cada palabra), la base de datos quedó conformada como se muestra en la siguiente tabla:

Palabra clave	**Reforma Energética en prensa**		
	LJ	**R**	**U**
pemex	113	113	70
energética	105	82	58
TOTAL	**218**	**195**	**128**

Palabra clave	**Influenza en prensa**		
	LJ	**R**	**U**
influenza	323	359	282
epidemia	143	135	147
TOTAL	**466**	**494**	**429**

FIGURA 4. Resultado total de la revisión hemerográfica de ambos temas en los tres periódicos

Al comparar los resultados de la hemeroteca con los obtenidos de *Google Noticias* se aprecia la notablemente inconsistencia entre ellos. En el tema energético la búsqueda de *La Jornada* encontró 199 notas electrónicas de las 218 que publicó la prensa, la consulta de *El Universal* mostró 264 por las 128 del impreso y en el caso de *Reforma* únicamente 11 de las 195.

Respecto del tema de la influenza, la consulta electrónica de *La Jornada* arrojó 684 de 466 notas impresas, *El Universal* 751 de 429 y en *Reforma* 49 de 494 notas.

Palabra clave	Reforma Energética en Google		
	LJ	R	U
pemex	101	6	114
energetica	98	5	150
TOTAL	**199**	**11**	**264**

Palabra clave	Influenza en Google		
	LJ	R	U
influenza	461	34	527
epidemia	223	15	224
TOTAL	**684**	**49**	**751**

FIGURA 5. Resultado total de la revisión en Google Noticias

Para entender mejor el comportamiento de *Google,* decidimos clasificar la información de la siguiente manera:

- Notas comunes.- aquellas notas encontradas por ambos medios: hemeroteca y motor de búsqueda
- Notas exceso.- notas que sólo aparecen en *Google*, por lo tanto son un exceso respecto de la base de datos elaborada a partir de la revisión hemerográfica

- Notas déficit-. notas que no aparecen en la consulta de *Google*, pero están en la base de notas de los periódicos

Las tres variables nos ayudaron a determinar la eficiencia del motor de búsqueda y a observar un panorama más detallado de los resultados para cada periódico.

TEMA	MEDIO	NOTAS (edición impresa)	Google	EXCESO	DÉFICIT	COMUNES
pemex/energética	LJ	218	199	51	70	148
	R	195	11	5	189	6
	U	128	264	198	62	66
influenza/epidemia	LJ	466	684	322	104	362
	R	494	49	34	479	15
	U	429	751	563	241	188

FIGURA 6. Total de las notas extraídas de la hemeroteca y *Google* de los tres diarios y los dos temas

La Jornada

De los tres diarios, *La Jornada* arrojó menos inconsistencias. De las 684 notas rescatadas en la edición impresa, la herramienta electrónica encontró 510; es decir que su rendimiento fue del 75%, asimismo la consulta en el buscador omitió 174; un déficit del 25%. El exceso fue superior ya que incrementó el número de notas a más del 50% respecto de la base original. Para este medio, la variable del exceso o las "inventadas por *Google*" fueron aquellas notas procedentes de los diarios regionales[viii], principalmente de Guerrero (60); Jalisco (49); Michoacán (63); Morelos (55); Oriente (120); San Luis (19) y otros (7).

Palabra clave	LJ	Google	Comunes	Exceso	Déficit
pemex	113	101	75	26	38
enegética	105	98	73	25	32
influenza	323	461	233	228	90
epidemia	143	223	129	94	14
TOTAL	**684**	**883**	**510**	**373**	**174**

Reforma

De las 689 notas publicadas en la edición impresa, *Google* encontró el 3% (21 notas), de ahí que la herramienta presentara un déficit del 97% respecto del grupo testigo y un exceso del 6% al reportar sólo 39 notas.

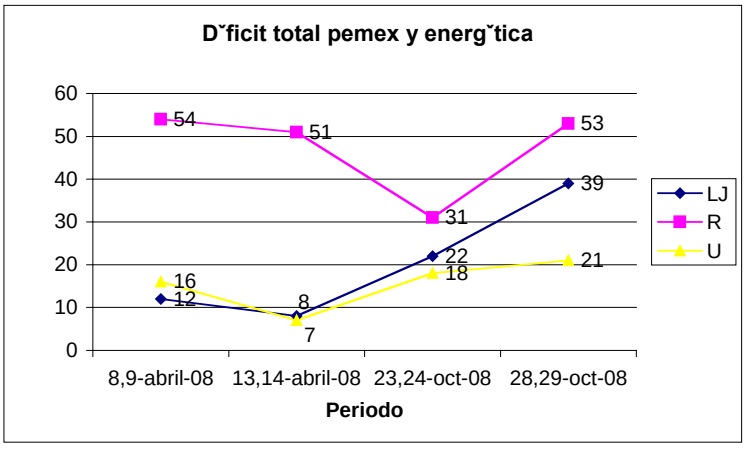
Palabra clave	R	Google	Comunes	Exceso	Déficit
pemex	113	6	3	3	110
energética	82	5	3	2	79
influenza	359	34	10	24	349
epidemia	135	15	5	10	130
TOTAL	**689**	**60**	**21**	**39**	**668**

El Universal

De los tres diarios, éste presentó mayor índice de exceso. Durante la revisión hemerográfica se reportaron 557 notas; en *Google* los valores se excedieron casi al doble con 1015 notas, de las cuales sólo 254 fueron comunes (46%). La búsqueda electrónica arrojó 761 notas de exceso y clasificamos la distribución de esta variable de la siguiente manera: 550 corresponden a las emisiones de *Minuto x Minuto*[ix], 3 son publicaciones del *El Gráfico*[x], 46 pertenecen a *El Universal* de Caracas, Venezuela y 163 a otros.

Palabra clave	U	Google	Exceso	Déficit	Comunes
pemex	70	114	76	32	38
energetica	58	150	122	30	28
influenza	282	527	407	162	120
Epidemia	147	224	156	79	68
TOTAL	**557**	**1015**	**761**	**303**	**254**

En un esfuerzo por detallar los datos obtenidos por *Google* presentamos gráficas por tema para mostrar las variables de exceso y déficit de los 16 experimentos.



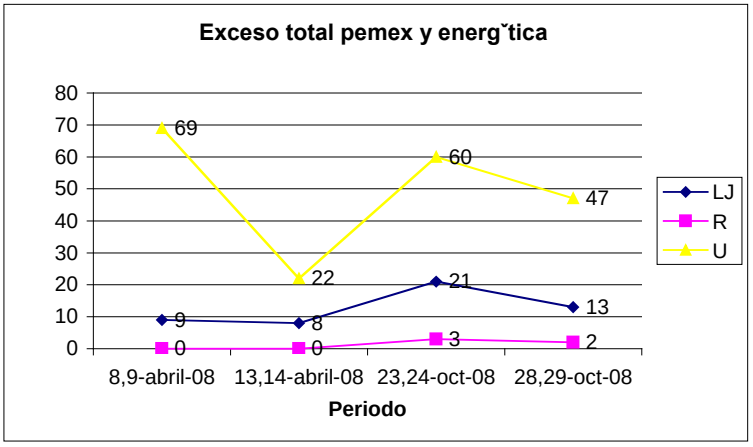
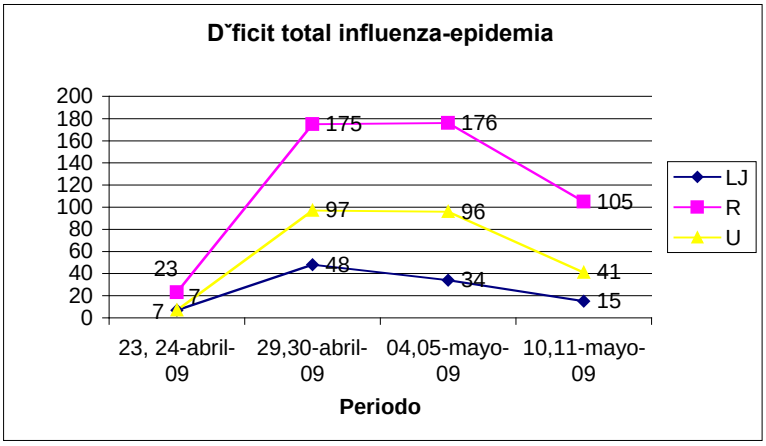


FIGURA 7. Déficit y exceso de Google en los tres diarios del tema energético



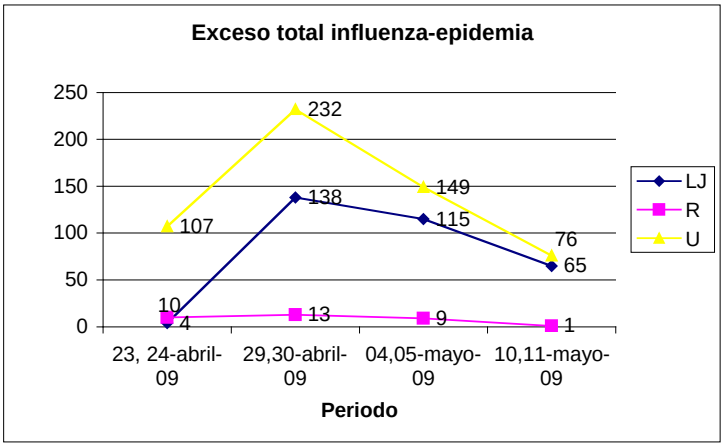


FIGURA 8. Déficit y exceso de Google en los tres diarios del tema de la influenza

Conclusión

Al ser un estudio puramente cuantitativo, encontramos que los valores finales obtenidos del experimento no proporcionan los argumentos válidos para explicar el proceder de *Google* durante las consultas. Lo que sí detectamos fue la existencia de indicadores externos

(formato en que los diarios indexan sus páginas) e internos (mecanismo del algoritmo de *Google*) causantes de las tendencias del déficit y exceso.

Los presentes resultados confirman, básicamente, que el comportamiento de *Google* depende de la amplitud de la cobertura del hecho noticioso, pues observamos que al incrementarse el número de notas aparecen mayores inconsistencias en el motor de búsqueda; por otro lado, cuando el número de notas se mantiene por debajo de las 100 se muestra cierta estabilidad. Con estos resultados enunciamos que *Google* no cumple con el rendimiento indispensable para compilar material hemerográfico sobre temas noticiosos de manera óptima.

También creemos que por las características y el tamaño del experimento, no es posible hacer una generalización sobre la eficiencia de *Google Noticias*; de hacerlo, sería necesario ampliar la muestra (como recomiendan los expertos en análisis estadístico). No obstante, *la Búsqueda Avanzada en el Archivo de Noticias de Google* puede brindar excelentes resultados a aquellos usuarios que necesitan una recopilación somera de notas sobre un tema específico para tener una noción del hecho y de la cobertura realizada por la prensa. Asimismo queda abierta la posibilidad de continuar con la investigación y mejorar los resultados obtenidos acerca de cómo los efectos de los indicadores externos e internos repercuten con los resultados electrónicos a través de un mayor número de consultas basadas en los parámetros de búsqueda establecidos.

Referencias

Islas, O y Benassini, C. (Coords.). (2005). *Internet, columna vertebral de la sociedad de la información* (67-91). México: ITESM, Campus Estado de México.

Lopez Yepes, J. (2004). *Diccionario Enciclopédico de Ciencias de la Documentación.* Madrid: Síntesis.

Milstein, S. Bierdorfer, J. D. y MacDonald, M. (2006). *Google: The missing manual.* Estados Unidos: O'Reilly Media.

Trejo Delarbre, Raúl. (1996). *La nueva alfrombra mágica: usos y mitos de Internet, la red de redes*. México: FUNDESCO.

Uyar, A. (2009). *Investigation of the accuracy of search engine hit counts*. En *Journal of Information Science*, Vol. 35 No. 4. Pp. 469-480.

[i] Cruz, Javier. *Cómo elegir (y comprender) las Fuentes en el periodismo de ciencia*. En Jornalismo e ciencia: uma perspectiva ibero-americana. Museo da Vida/ Casa de Oswaldo Cruz/ Fiocruz, 2010. p. 46.

[ii] Ahmet Uyar. *Investigation of the accuracy of search engine hit counts*. Journal of Information Science, 2009.

[iii] José Lopez Yepes, *Diccionario Enciclopédico de Ciencias de la Documentación*. Madrid, 2004.

[iv] Ahmet Uyar. *Investigation of the accuracy of search engine hit counts*. Journal of Information Science, 2009. pp. 469.

[v] La mayoría de los motores de búsqueda ofrecen esta opción, en la cual el usuario introduce una consulta y el rastreador realiza una búsqueda general, sin especificaciones ni parámetros que delimiten los resultados.

[vi] Google selecciona artículos de miles de fuentes de noticias en línea, y luego los presenta por tema y por categoría. Google Noticias utiliza los sofisticados algoritmos de computadora para agrupar y clasificar las noticias. La ventaja de este sistema es que Google puede recopilar historias mucho más rápido que la mayoría de los servicios de agregación de noticias. Sarah Milstein, J. D. Bierdorfer y Matthew MacDonald, *Google: The missing manual*, Estados Unidos, O'Reilly Media, 2006, 2a edición, pp. 91 y 92.

[vii] *Altavista* al igual que *Google* ofrece campos específicos para la búsqueda especializada requerida, sin embargo al someterla a las pruebas de consulta nos encontramos que sólo funciona con algunos diarios y en otros casos nos traslada al buscador de *Yahoo*.

[viii] La Jornada cuenta con una serie de periódicos que cubren áreas regionales dentro de la República Mexicana, estás son*: La Jornada Aguascalientes, La Jornada Guerrero, La Jornada Jalisco, La Jornada*

Michoacán, La Jornada Morelos, La Jornada Oriente, La Jornada San Luis, La Jornada Veracruz, La Jornada Zacatecas

[ix] Sección de noticias breves publicadas exclusivamente en versión electrónica, abarcan todas las secciones y son actualizadas cada 4 minutos en línea

[x] Periódico editado por *El Universal*, de formato sencillo y con menor cobertura de noticias

Javier Crúz Mena. Fisico de la UNAM, con estudios de posgrado en las universidades de Princeton (ingeniería química) y Brown (matemáticas). Ha ejercido el periodismo de ciencia en varios medios de comunicación desde 1994. Actualmente es colaborador en el *Noticias MVS* con Carmen Aristegui, dirige el programa la *Escencia de la ciencia* en el Instituto Mexicano de la Radio y es editor de la Unidad de periodismo de ciencia en la Dirección de General de Divulgación de Ciencia en la UNAM.

María Keninseb García Rojo. Es egresada de la carrera de Ciencias de la Comunicación de la Facultad de Ciencias Políticas y Sociales de la UNAM, generación 2007-2011. De septiembre de 2008 a agosto de 2010 fue becaria de la Dirección General de Divulgación de la Ciencia en el Departamento de Noticias y Documentación. Actualmente trabaja en el Departamento de Prensa y Difusión del Instituto de Investigaciones Biomédicas de la UNAM.

Isela Alvarado Cruz. Es egresada de la carrera de Ciencias de la Comunicación de la Facultad de Estudios Superiores Acatlán de la UNAM (2007). Actualmente es tesista y realiza prácticas profesionales en la Unidad de periodismo de ciencia de la Dirección General de Divulgación de la Ciencia, UNAM.